International Journal of Information and Computer Science
IJICS Volume 1, Issue 2, May 2012 PP. 46-51
ISSN(online) 2161-5381
ISSN(print) 2161-6450
www.iji-cs.org

Study on the Protection Method of Data Privacy Based on Cloud Storage

Zhang Shaomin, Li Xiaoqiang, Wang Baoyi
North China Electric Power University, dept. of Computer, Hebei Baoding 071003, China

(Abstract) In this paper, we first analyze the concept, model, infrastructure and security problems of cloud storage. Then aiming at the feature of distributed file system,we systematically design the corresponding methods of data privacy protection. A method based on distributed file system to protect data privacy in cloud storage is proposed. This method makes full use of the characteristic of master-slave structure system in the distributed file system, converting the traditional direct encryption of data files to the encryption of metadata in server, and thus shortens the encryption time effectively in privacy protection. Experiments show that our method has a better performance on data privacy protection in great different sizes of data files, which is especially suitable for great different data sizes and high time efficiency requirement environments, effectively making up the deficiency in encryption time of the traditional protection method, promoting the development and application of cloud storage.

Key Words: Cloud Storage; Data Privacy; Metadata

1. INTRODUCTION

Cloud storage is a new extension and developed from the concept of cloud computing. Cloud storage is a system which uses a large number of different types of storage devices working together through network by the technology of the cluster, grid or the distributed file system to provide data storage and access services. Cloud computing system transforms into a cloud storage system when the core operation and function are handling and storing large amounts of data and it need to configure a large number of storage devices. So cloud storage is a cloud computing system with its core function is data storage and management.

The survey of IDC (Internet Data Center) shows that cloud computing services will be rapid growing in next five years and is expected to reach the market size of $42 billion in 2012. At present, the enterprises using cloud computing are become popular and growing year by year. It is estimated that the invest in cloud computing services will account for 25% of all IT costs in 2012, and even to one-third of the total IT costs in 2013. However, the development of cloud computing faces with many critical issues and the security issues is the most important ones. With the increasing popularity of cloud computing, the important of the security issues trend rise gradually, and have become an important factor restricting its development.

The Gartner's survey shows that more than 70 percent of respondents CTO will not bring in cloud computing recently because of the concern about data security and privacy. Google's user files leakage event and Amazon Simple Storage Service interrupted accidents exacerbate the concerns of people [3]. At present, governments and security organizations and others take cloud computing security very serious and put a great deal of resources researching and setting the standards of cloud computing security, in order to promote the application and development of cloud computing.

Cloud storage platform involves many security issues. The primary problems include infrastructure safety, virtualization security, data security and application service security and so on. And the data security is the single most important one because user has little control on their data in cloud storage. They don't know the position of their data, which servers handle their data, what network they transmitted by, and even where they stored in [4]. The high flexibility and expansibility of cloud computing not only make cloud computing attracting but also lead to the difficulty in predicting and guarantying user's data security. Data security is mainly to provide security guarantees for the data stored in cloud storage. Data security in the cloud storage includes data confidentiality, privacy, availability, integrity, authenticity, authorization, authentication and non-repudiation and so on.

According to the cloud security model of Cloud Security Alliance [5-6], the core technologies of cloud security are cryptographic and reinforcement techniques. Reinforcement technique takes a large number of cryptographic techniques to provide user with a trusted secure cloud. In the paper, we propose a method based on distributed file system to protect data privacy in cloud storage. It will be effective in protecting the privacy of the data stored in the cloud.

2. THE MODEL OF CLOUD STORAGE

2.1 The Structural Model of the Cloud Storage

Cloud storage platform structure can be divided into four layers, namely the storage layer, basic management layer, application interface layer and access layer, just as shown in Figure 1 below.

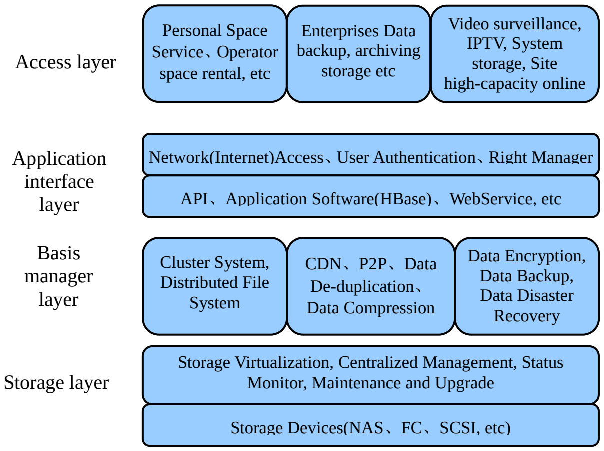

Access layer
Personal Space Service、Operator space rental, etc
Enterprises Data backup, archiving storage etc
Video surveillance, IPTV, System storage, Site high-capacity online
Application interface layer
Network(Internet)Access、User Authentication、Right Manager
API、Application Software(HBase)、WebService, etc
Basis manager layer
Cluster System, Distributed File System
CDN、P2P、Data De-duplication、Data Compression
Data Encryption, Data Backup, Data Disaster Recovery
Storage layer
Storage Virtualization, Centralized Management, Status Monitor, Maintenance and Upgrade
Storage Devices(NAS、FC、SCSI, etc)

Fig. 1 Cloud storage model

(1)The storage layer

The storage layer is the most basic part of the cloud storage. Storage devices include all types of different equipments, such as Fabre Channel storage devices, NAS (network storage devices), IP (Internet Protocol) storage devices, etc. A large number of these storage devices are located in different geographical locations, and connected with each other by network or Fabre Channel (FC). There is a unified storage device system to manage the storage device virtualization logic, the multi-link redundancy and the hardware status monitoring, fault maintenance.

(2)The basis management layer

The basis of management is the most core part of the cloud storage system. The basis management layer achieves interoperability among the multiple storage devices in the cloud storage by the technology of cluster, grid computing and distributed file system. The multiple storage devices can provide the same services, and provide better properties of data access. Content distribution system and the data encryption technology can ensure data will not be accessed by unauthorized users. At the same time, by using all kinds of data backup and tolerant measures system can ensure the data has high integrity and stability, almost never lost, and thus ensure its own security and stability.

(3)The application interface layer

The application interface layer is flexible. According to the actual business types, different cloud storage providers may develop different application service interfaces and provide different application services, such as video surveillance application platform, IPTV, video-on-demand application platform, network hard disk application platform, remote data backup application platform and so on.

(4)The access layer

Any authorized user can login cloud storage platform through standard public application interface, enjoys cloud storage services. Different cloud storage enterprises provide different kinds of access types and methods.

2.2 The Infrastructure of the Cloud Storage

The distributed file system is one of the most important parts in cloud storage. The distributed file system likes a huge pool of storage resources and a large number of low-cost storage devices integrated as a whole to provide a unified storage services for upper. Most the distributed file system consists of three major components (as shown in Figure 2), including the distributed file System client, meta-data server, and block-data server. Distributed file system can be accessed by multiple clients; block-data servers (datanode) are generally composed of many servers; meta-data server (namenode) is depending on the design of distributed file system, may be a server or multiple servers [8].

Generally speaking, meta-data sever is responsible for allocating and recording the position of every block for all files in block-data servers and storing their metadata information in the distributed file system. The distributed file system client will divide file and send the blocks to the special block-data servers which is designated by meta-data server, or query meta-data server to obtain the location information for data block then read data from the corresponding block-data servers. In a word, the block-data servers are charge for the block file storage.

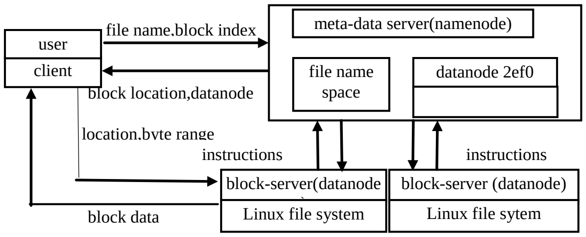


Fig. 2 The structure of distributed file system

3. DATA PRIVACY PTOTECTION DESIGNED FOR CLOUD STORAGE

3.1.The Encryption Storage Method Based on Client Data

The encryption storage method based on client data is conventional, traditional encryption scheme. The main idea is to encrypt the data before storing to protect user data privacy effectively. When storing data, user first encrypts the data then stores the data in the cloud. At data access, user reads data directly from the cloud then decrypt the data.

(1) Data storage

Firstly, user encrypts the data then put them in the cloud. The specific steps are shown in Figure 3.

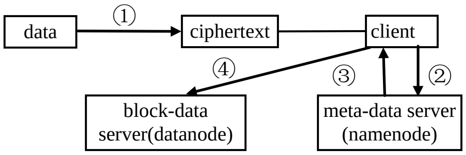


Fig. 3 The storage steps of encryption storage method based on client data

step 1：The client obtains the cipher text by encrypting the data to be stored in the cloud(Figure ①);

step 2：The client interacts with meta-data server, sends the meta-information(size, etc) of cipher text to meta-data server(Figure②). Then meta-data server allocates storage space and return these metadata back to the client(Figure③);

step 3：With the metadata from meta-data server, the client will connect the corresponding block-data servers, establish reliable channel and send the data to block-data servers to storage (Figure④).

(2) Data access

The client reads data directly from the cloud, and then decrypts the cipher text from cloud. The specific steps are shown in Figure 4.

Fig. 4 The access steps of encryption storage method based on client data

step 1：The client sends a request to meta-data server with the directory of the distribute file system for the metadata to the corresponding file(Figure①). The meta-data server queries the file naming system for the needed metadata and returns them to client (Figure②).

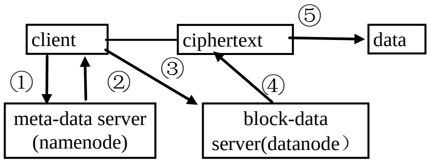
step 2：According to the metadata from meta-data server, the client communicates with the corresponding block-data servers(Figure③) and acquires cipher text from block-data servers(Figure④).

step 3：After all data blocks has been received, the client encrypts cipher text to the request data(Figure⑤).

(3) The analysis of the privacy protection

In the encryption storage method based on client data, data in the cloud is encrypted. The privacy of data is achieved through the encryption algorithm. Encryption algorithms have been able to fully guarantee their privacy in modern cryptography. Without the right key information, even if someone gets the cipher data, it would not leak user's data. As a result, the encryption storage method based on client data is able to protect the privacy of user's data in the cloud. But if the data is much bigger, it costs such long time encrypting the data, which lead to a performance bottleneck and make a very bad influence on the efficiency in reading and writing.

3.2.The Encryption Storage Method Based on Server Metadata



In the distributed file system, the storage of the data information (metadata) and the specific data are separately. Data storage information is stored in the meta-data server, and the specific data are stored in the block-data server. When you want to read a file, you must firstly read metadata from the meta-data server to acquire the data location information, and then obtain specific data from block-data servers. So long as the well protection of metadata in meta-data server will be able to protect data privacy effectively. In the cloud storage environment, no one knows which one or which several block-data servers the data are stored in except for meta-data server. When the client wants to access the data, he must acquire the metadata from meta-data server at first, then connect corresponding block-data severs, complete reading or writing data. The metadata in the meta-data server are the key to the cloud storage. It is great difficult to obtain metadata by traversing the block data server. On one hand, the number of the block-data server is large and lacking information, on other hand without the necessary metadata such as size, format, encoding, etc, it is hard to validate the boundary of the data in the cloud storage. It can be said that encryption the metadata in the meta-data server will also be able to protect the privacy of data in cloud storage, especially relatively large data. Because they may be divided into smaller parts and stored in different block-data servers. Owing to the very little metadata in the meta-data server, the encryption and decryption efficiency is very high.

(1)Storage data

When user storing data, the meta-data server allocates storage space firstly and then hides the meta-information by encryption. User's data are stored in the block-data server normally。The specific steps are shown in Figure 5.

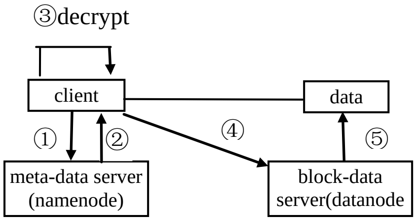


Fig.5 The storage steps of encryption storage method based on server metadata

step 1：The client communicates with the meta-data server, and sends the information of the storage data and the user's key to meta-data server. The meta-data server allocates storage space in the distributed file system and returns the metadata to client(Figure①②);

step 2：In the meta-data server, meta-data server hides the meta-information acquired from step 1 by encryption with the client's key(Figure③);

step 3：With the metadata from meta-data server, the client is able to connect the corresponding block-data servers, establish reliable channel and send data to block-data servers for storage(Figure④).

(1) Data access

When reading, after acquiring cipher metadata from meta-data server, the client need to decrypt it with user's secret key to gain the real metadata. Only after that, user is able to connect and acquire data from block-data servers. The specific steps are shown in Figure 6.

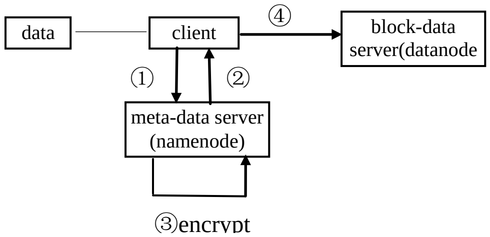


Fig. 6 The access steps of encryption storage method based on server metadata

step 1：The client sends a request to meta-data server with the directory of the distribute file system for the metadata to the corresponding file(Figure①). The meta-data server queries the file naming system for the corresponding metadata and returns them to client(Figure②). At this moment, the metadata are encrypted.

step 2：The client decrypts cipher metadata with user's secret key to get the real metadata (Figure③).

step 3: The client communicates with the corresponding block-data servers and acquires data from block-data servers, so as to obtain all data or file(Figure④⑤).

(3) The analysis of the privacy protection

In the encryption storage method based on server metadata, metadata is encrypted in the cloud. The privacy of data is achieved by the encryption and hidden of key information. Without the right key information, even if small data, no one can acquire metadata, so as to the user's data in cloud. In this method, it only need to encrypt the metadata in the meta-data server, while the metadata in the meta-data server is little and memory-resident. So it costs little time to encrypt the data, making a very good influent on the efficiency of reading and writing.

3.3.The Running Performance of the Two Protection Methods And Their Applications

(1) The analysis of running efficiency

According to the encryption process of the encryption algorithm, we can get below formula:

$T = F(AM, S)$ ①

Among them, T represents the total time of encryption;

AM represents the encryption algorithm;

S represents the size of the data file which is to be encrypted.

As for formula ①, when AM is determined, the total time of encryption T and the size of the data file S are linearly dependent. That is to say, the bigger size of data file, the more time the encryption requires. Figure 7 below shows the relationship between the encryption time of AES and the size of data file with the key length of 256 bits.

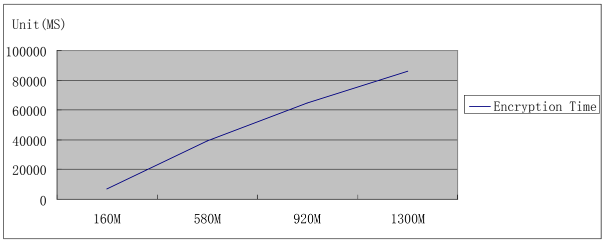


Fig.7 The the relationship between the encryption time of AES and the size of data file

The encryption method based on server metadata changes the direct encryption of data file into the encryption of metadata in meta-server by the special characteristic of the distributed file system (HDFS). And the size of metadata is far less than the data file in HDFS. The total time of the encryption method based on server metadata is far less than the method encrypting data directly on the basis of the linear characteristics of the encryption time. Hence, the encryption method based on server metadata has a better performance in data privacy protection.

(2) The applications of the two protection methods

Both the encryption storage method based on client data and the encryption storage method based on server metadata use encryption to protect the data privacy in cloud. The only difference is the encryption content. The first method need to encrypt all data, it will be more secure but cost longer time while the second program only need to encrypt metadata in the cloud storage, it will be less secure but the encryption time is very short and has a higher efficiency on data privacy protection.

As for the encryption storage method based on client data, although its safety is higher but its encryption time limitation restricts its application. As to the encryption storage method based on server metadata, because the metadata are little, whether big or small data, it has a better efficiency in data privacy protection. The encryption storage method based on client data is suitable for the occasion of the small and sensitive data file, such as account information, etc. And the encryption storage method based on server metadata can be used in the big data and relatively insensitive environment, and the bigger data file the more safe.

4. EXPERIMENT AND ANALYSIS

In order to verify the high efficiency of the encryption storage method based on server metadata, we have built a small hadoop cluster. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage [9].

In this experiment, the cluster has a total of three machines. All the machines are of the same configuration, each having main memory size of 2G, hard disk of size 160G and the operating system is window7. Besides, each machine has installed the jdk-6u26-windows-i586, hadoop-0.20.2 and Cygwin.

The level of data privacy protection efficiency is mainly reflected in the time consumption. Therefore we set a storage time consumption of data file as the efficiency benchmarks, and use the weighted average method to calculate the storage time in different data files. As for encryption algorithm, we adopt AES (Advanced Encryption Standard) with the key length of 256 bits.

(1)The efficiency test of data privacy protection in different sizes of data files

We use different sizes of data files to test the cost time in data storage. Table 1 below shows the results in the form of time consumption for different sizes of files and for different approaches of encryption.

Tab. 1 Comparison of different protection methods to storage time-consuming

NO.	File size（M）	No Protection（S）	Encryption Storage based on meta-data（S）	Encryption Storage based on client data（S）
1	10	1.2	1.4	1.9
2	160	11.1	11.4	18.3
3	580	56.4	57.1	107
4	930	86.2	87.1	154.1
5	1300	137	138.8	240.9

In order to display the comparison results in a more intuitive way, we use a line chart below. The vertical axis is time, with the unit of second. The sizes of data files are represented along the x-axis.

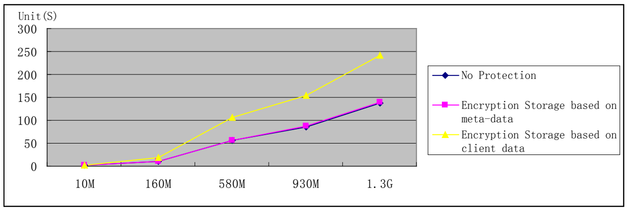


Fig. 7 The comparison chart of different encryption method in storage time-consuming

Through the experimental comparison, we can find that the encryption storage method based on server metadata has a better efficiency in data privacy protection and the gap rises with the increasing data file. Also we can find that the efficiency of the encryption storage method based on server metadata is close to no protection. This is mainly due to the block size in the distributed file system which is 64M and is much greater than the traditional file system. So it needs less metadata when data are stored in distributed file system. For example, when storing a 10G file in HDFS, it only needs 160 metadata and the metadata is little. Hence the privacy protection efficiency is very high. When the file and the cluster become bigger, the data privacy efficiency becomes higher. In addition, we can find that the gap among three methods is a little bit different when the data file is less than 10M from the figure. It is necessary to further verify the difference among them in the case of small data files.

(2)The efficiency test of data privacy protection in small data file

In the experiment we used a thousand of 10KB data files continuously stored in HDFS. The time spending in storage of the no protection method, the encryption storage method based on server metadata and the encryption storage method based on client data are shown in Figure 8.

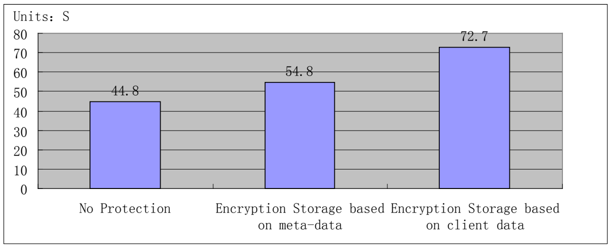


Fig. 8 The comparison chart of different protection method in continuous storage time consumption for small data file

According to the experimental comparison, we can find that even in the case of small file, the encryption storage method based on server metadata still has a higher efficiency than the encryption storage method based on client data. This is mainly because the metadata is very small and memory-resident, and the encryption time is very short and which lead to high protection efficiency. As a result, the encryption storage method based on server metadata still has a better performance in the case of the small file.

5. SUMMARY

In this paper, we researched the data privacy protection in the cloud storage. We proposed a method based on distributed file system to protect data privacy in cloud storage. Compared to the traditional methods, it has a higher efficiency in great different sizes of data. The experiences in small Hadoop cluster prove the superiority of the method in the protection of data privacy. Considering the problems of data retrieval and key management, the application also needs to consider the problem of the encryption algorithm and the following work is encryption algorithm.

6. REFERENCES

[1] http://www.cloudcomputing-china.cn/Article/luilan/201003/564.html

[2] Liu peng. Cloud Computing(Second Edition)[M]. Beijing： Publishing House of Electronics Industry. 2011

[3] FENG Deng-guo, ZHANG Min, ZHANG Yan,etal. Study on Cloud Computing Security. Journal of Software, 2011,22(1):71-83.

[4] LIN Zhao-ji, FU Xiong, WANG Ru-chuan,etal. Research on Security Challenges in Cloud Computing. Information Research. 2011,37(2):1-4.

[5] Cloud Security Alliance: Security Guidance for Critical Areas of Focus in Cloud Computing V2.1. https://cloudsecurityalliance.org/guidance/csaguide-cn.v2.1.pdf

[6] Wang Huibo. Security storage and cloud storage security. Information security and confidentiality of communications [J].2012.12: 18-19

[7] Liu Bei, Tang Bin. Principles and Development Trend of Cloud Storage. Science & Technology Information[J]. 2010.05: 470-471.

[8] HOU Qin-hua, WU Yong-wei, ZHENG Weimin, etal. A Method on Protection of User Data Privacy in Cloud Storage Platform. Journal of Computer Research and Development. 2011,48(7):1146-1154.

[9] What is Apache Hadoop. http://hadoop.apache.org/

[10] Guo Chunmei, Bi Xueyao, Yang Fan. Cloud Computing Security Technology Research and Trends. Information Network Security. 2010, 4: 16-17

[11]Somani, U; Lakhani, K.; Mundra, M.. Implementing digital signature with RSA encryption algorithm to enhance the Data Security of cloud in Cloud Computing[J]. Parallel Distributed and Grid Computing (PDGC), 2010 1st International Conference on . 2010 , Page(s): 211 – 216

[12]Sabahi, F. Cloud computing security threats and responses[J]. Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on . 2011 , Page(s): 245 – 249

[13] Sengupta, S.; Kaulgud, V.; Sharma, V.S. Cloud Computing Security--Trends and Research Directions[J]. Services (SERVICES) , 2011 IEEE World Congress on. 2011 , Page(s): 524 - 531